# Stock market predictions and observations

By:
Nadav Schneider
Dolev Gelbert
Or Elazar

# Purpose of the project

Making predictions and observations about stock market behavior stocks prices and trends.

By collecting historic and up to date data about different stocks throughout the years, we will try to predict future stocks prices and behavior.

We will develop models to help us predict the following:

- The price of a specific stock in the near future.
- If a market is to go up or down the following month
- Is a stock a good long term- medium term  (2 years-3 years) investment.

# Sources

Wikipedia :

https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

https://en.wikipedia.org/wiki/Nasdaq-100

https://en.wikipedia.org/wiki/TA-125_Index

Wallstreetzen:

https://www.wallstreetzen.com/stock-screener/stock-forecast

Yhaoo Finance API

# Data Collection Process

We combined the use of crawlers and api to create different data frames, from which we explored viable prediction problems, tested different methods and models, made observations and trained our machine learning models.

At first we collected lists of stocks names and ticker symbols of stocks bundled in major indexes by crawlers.

Using Yhaoo finance API we retrieved historical and current data about those stocks.

We also collected data of a wider range of stocks (not necessarily of big companies) using a crawler and collected experts recommendations about those stocks to use as a reference point.

| | | |
|---|---|---|
| **0** | 3M | MMM |
| **1** | A. O. Smith | AOS |
| **2** | Abbott | ABT |
| **3** | AbbVie | ABBV |
| **4** | Abiomed | ABMD |
| **...** | ... | ... |
| **727** | Levinstein Properties Ltd. | LVPR |
| **728** | Hadera Paper Ltd. | HAP |
| **729** | FMS Enterprises Migun Ltd. | FBRT |
| **730** | Doral Group Renewable Energy Resources Ltd. | DORL |
| **731** | Gilat Satellite Networks Ltd. | GILT |

]:

| | Open | High | Low | Close | Adj Close | Volume | increased |
|---|---|---|---|---|---|---|---|
| **Date** | | | | | | | |
| **1972-02-01** | 0.000000 | 107.160004 | 103.099998 | 106.570000 | 106.570000 | 376260000 | 1 |
| **1972-03-01** | 0.000000 | 109.750000 | 105.860001 | 107.199997 | 107.199997 | 403650000 | 1 |
| **1972-04-01** | 0.000000 | 111.110001 | 106.180000 | 107.669998 | 107.669998 | 367990000 | 1 |
| **1972-05-01** | 0.000000 | 111.480003 | 103.830002 | 109.529999 | 109.529999 | 335850000 | 1 |
| **1972-06-01** | 0.000000 | 110.510002 | 105.940002 | 107.139999 | 107.139999 | 314510000 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **2020-08-01** | 3288.260010 | 3514.770020 | 3284.530029 | 3500.310059 | 3500.310059 | 84402300000 | 1 |
| **2020-09-01** | 3507.439941 | 3588.110107 | 3209.449951 | 3363.000000 | 3363.000000 | 92084120000 | 0 |
| **2020-10-01** | 3385.870117 | 3549.850098 | 3233.939941 | 3269.959961 | 3269.959961 | 89737600000 | 0 |
| **2020-11-01** | 3296.199951 | 3645.989990 | 3279.739990 | 3621.629883 | 3621.629883 | 100977880000 | 1 |
| **2020-12-01** | 3645.870117 | 3760.199951 | 3633.399902 | 3756.070068 | 3756.070068 | 96056410000 | 1 |

587 rows × 7 columns

]:

| | AAPL | ABC | ABT | ADI | ADM | ADP | ADSK | AEP | AES | AIG | ... | VTRS | VZ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Date** | | | | | | | | | | | | | | |
| **1999-12-01** | 0.917969 | 3.796875 | 16.300713 | 46.500000 | 10.997732 | 42.726330 | 8.437500 | 32.125000 | 37.375000 | 1441.666626 | ... | 11.194444 | 55.364555 | 2 |
| **2000-01-01** | 0.926339 | 4.531250 | 14.617335 | 46.750000 | 10.657596 | 37.620979 | 7.640625 | 33.500000 | 40.062500 | 1391.666626 | ... | 11.833333 | 55.701801 | 3 |
| **2000-02-01** | 1.023438 | 3.640625 | 14.813729 | 78.625000 | 9.126984 | 34.547855 | 11.171875 | 28.125000 | 41.906250 | 1179.166626 | ... | 10.222222 | 44.010609 | 4 |
| **2000-03-01** | 1.212612 | 3.750000 | 15.795700 | 80.500000 | 9.353741 | 38.265343 | 11.375000 | 29.812500 | 39.375000 | 1460.000000 | ... | 12.222222 | 54.971104 | 4 |
| **2000-04-01** | 1.107701 | 5.000000 | 17.254627 | 76.812500 | 9.013605 | 42.676762 | 9.593750 | 36.625000 | 44.968750 | 1462.500000 | ... | 12.611111 | 53.959366 | 4 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **2021-08-01** | 151.830002 | 122.209999 | 126.370003 | 162.949997 | 60.000000 | 209.039993 | 310.089996 | 89.570000 | 23.870001 | 54.560001 | ... | 14.630000 | 55.000000 | 41 |
| **2021-09-01** | 141.500000 | 119.449997 | 118.129997 | 167.479996 | 60.009998 | 199.919998 | 285.170013 | 81.180000 | 22.830000 | 54.889999 | ... | 13.550000 | 54.009998 | 35 |
| **2021-10-01** | 149.800003 | 122.019997 | 128.889999 | 173.490005 | 64.239998 | 224.490005 | 317.609985 | 84.709999 | 25.129999 | 59.090000 | ... | 13.350000 | 52.990002 | 36 |
| **2021-11-01** | 165.300003 | 115.750000 | 125.769997 | 180.250000 | 62.209999 | 230.889999 | 254.190002 | 81.050003 | 23.379999 | 52.599998 | ... | 12.310000 | 50.270000 | 32 |
| **2021-12-01** | 177.570007 | 132.889999 | 140.740005 | 175.770004 | 67.589996 | 246.580002 | 281.190002 | 88.970001 | 24.299999 | 56.860001 | ... | 13.530000 | 51.959999 | 37 |

# Data cleaning and sorting

Each machine learning model we created imposed different needs, for each of those we manipulated our data to fit our needs.

At first, not all of the stocks we were trying to collect data about had the relevant or sufficient information, on top of that data from some specific dates wasnt available. We removed dates and stocks that didnt contain sufficient data for our research.

For making an assessment If a market is to go up or down the following month, we had to manipulate our data some more, we added columns containing labels used a more detailed version of our dataframe.

# Data manipulating

For creating the assessment Is a stock a good long term- medium term (2 years-3 years) investment,We had to manipulate all our data, accumulate stocks prices, compute monthly change in stocks and average changes, we created a focused df out of our main df for the purpose of creating this model.We then normalized and labeled all of the data.

For the logistic regression model we had

to extract certain stock info from our

df insert it into new df label all of our data and

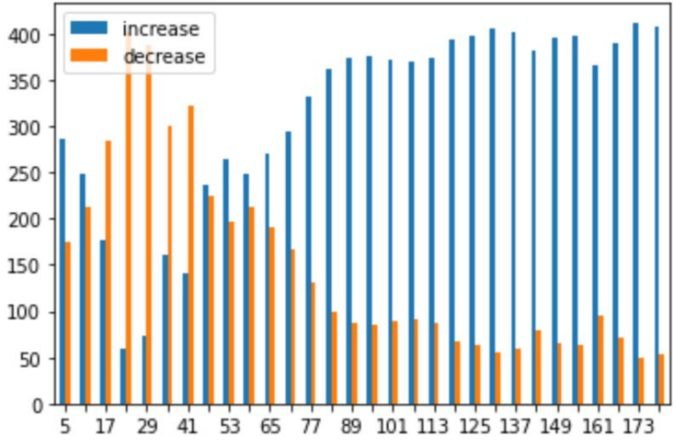
Convert categorical variable into dummy/indicator variables.

| | change | last_price | label |
|---|---|---|---|
| **AAPL** | -0.075878 | 0.031505 | Strong Buy |
| **MSFT** | -0.037939 | 0.024409 | Strong Buy |
| **GOOGL** | 0.203997 | 0.065108 | Strong Buy |
| **AMZN** | -0.392740 | 0.096976 | Strong Buy |
| **TSM** | -0.061825 | 0.026200 | Buy |
| **...** | ... | ... | ... |
| **ARW** | 0.089382 | 0.038809 | Hold |
| **GGB** | 0.674426 | 0.007119 | Buy |
| **WF** | 0.031507 | 0.016615 | Hold |
| **CLF** | 0.340034 | 0.126557 | Buy |
| **ALV** | 0.540749 | 0.011979 | Buy |

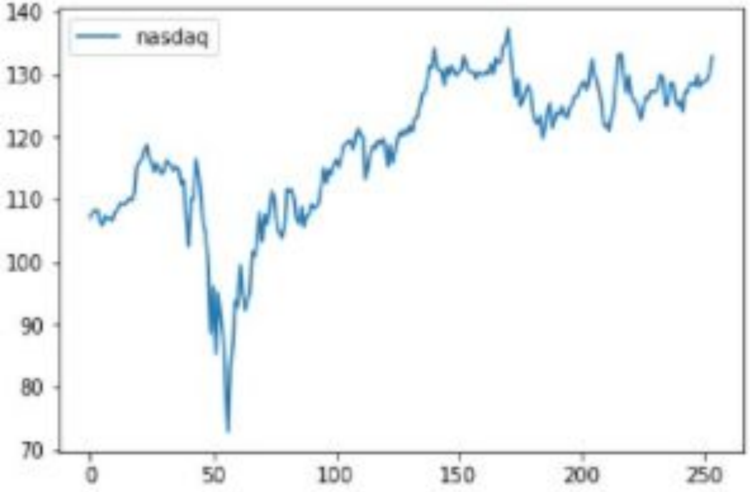
665 rows × 3 columns

# EDA - visualization and conclusions

the graph below visualize the ratio between amounts of stocks the increase and decrease their value over the years in increasing time periods, as we can see, in long time periods most stocks will increase their values, so we decided to create a model able to identifying those stocks you should stay away from and not get in for a long term investment.

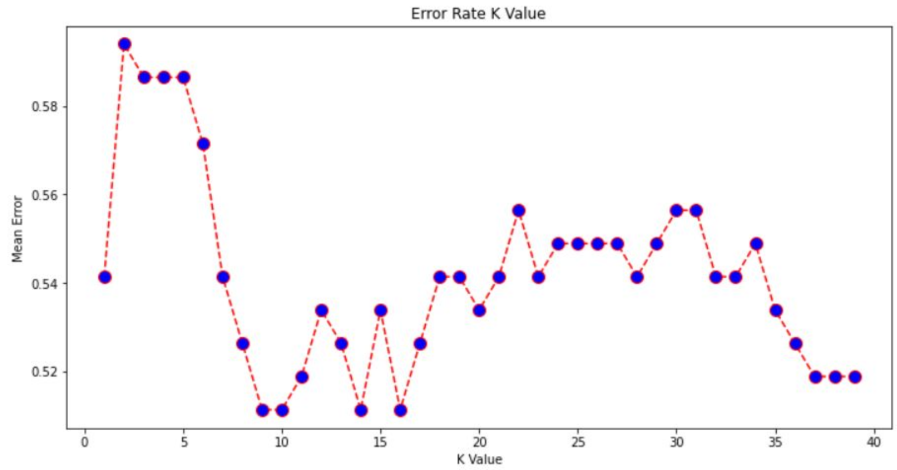


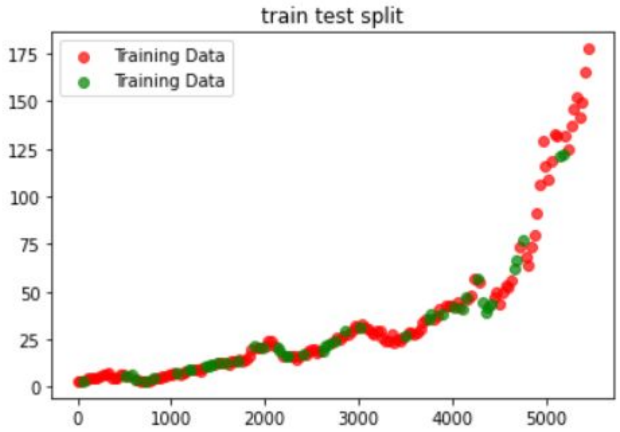
The graph below shows the nasdaq market Change in price over the years, as you can see the grow when looking long term is almost Linear, so we decided to create a linear model that will Predict future stock prices.Aside from these graph we used graph as a tool to visualize and help Us implementing and testing our machine learning models.

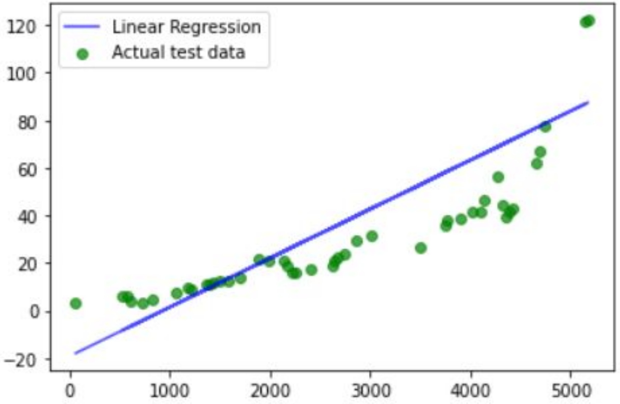


# Graph to determine the best k value to use in KNN model

```
0.48872180451127817
```




Estimating our linear model efficiency

# Machine Learning Models

We used 3 machine learning models in our project:

- LogisticRegrssion
- LinearRegression
- KNN

# Logistic Regression

We used logistic regression model to determine if a market is to go up or down the following month, by labeling and analyzing data from 1985 till today about NASDAQ. We included opening price, closing price, low price, high price, and volum from 35 years in 1 month intervals.

Eventually we came up with a model that is able to predict with a 0.7 certainty a markets behavior for the next month.

```
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
```

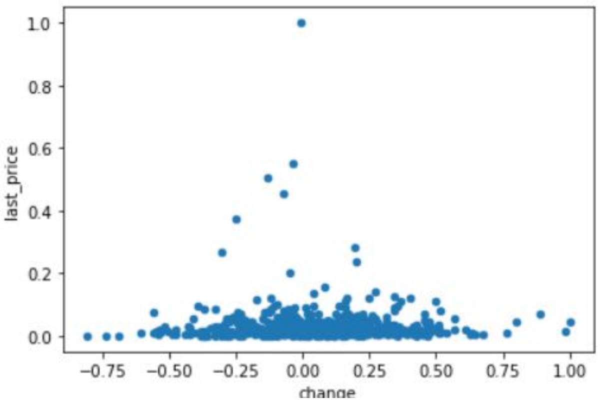
```
LogisticRegression()
```

```
log_reg.score(X_test,y_test)
```

```
0.6944444444444444
```

# KNN

We used knn model to predict if a stock is a solid long-medium term investment. By accumulating an labeling data from over 650 stocks on the course of over 3000 different sampling dates in the past we were able to determine in an underwhelming accuracy of 0.48 if a stock will rise significantly over the next 3 years.



# Linear Regression

We used linear regression model to predict a specific stocks price. We tested the data on different stock with different time intervals and found out the in monthly intervals for the course of a few years most stocks grow in an almost linear manner.

```
predict_stock_price('AAPL',stocks_data_df)
```